



Munich Center for Machine Learning



Human-Centric Methods in NLP

MCML-Lamarr Workshop

Michael A. Hedderich

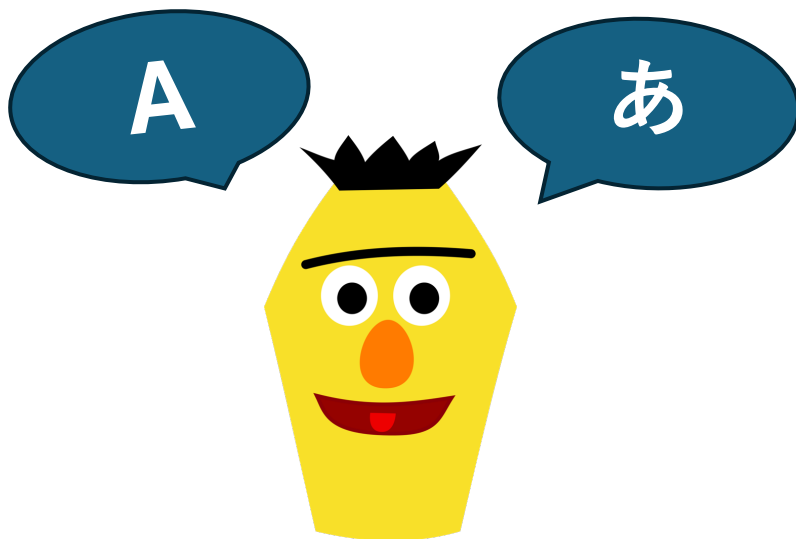
25.09.2025

My Personal Path to Human-Centric NLP



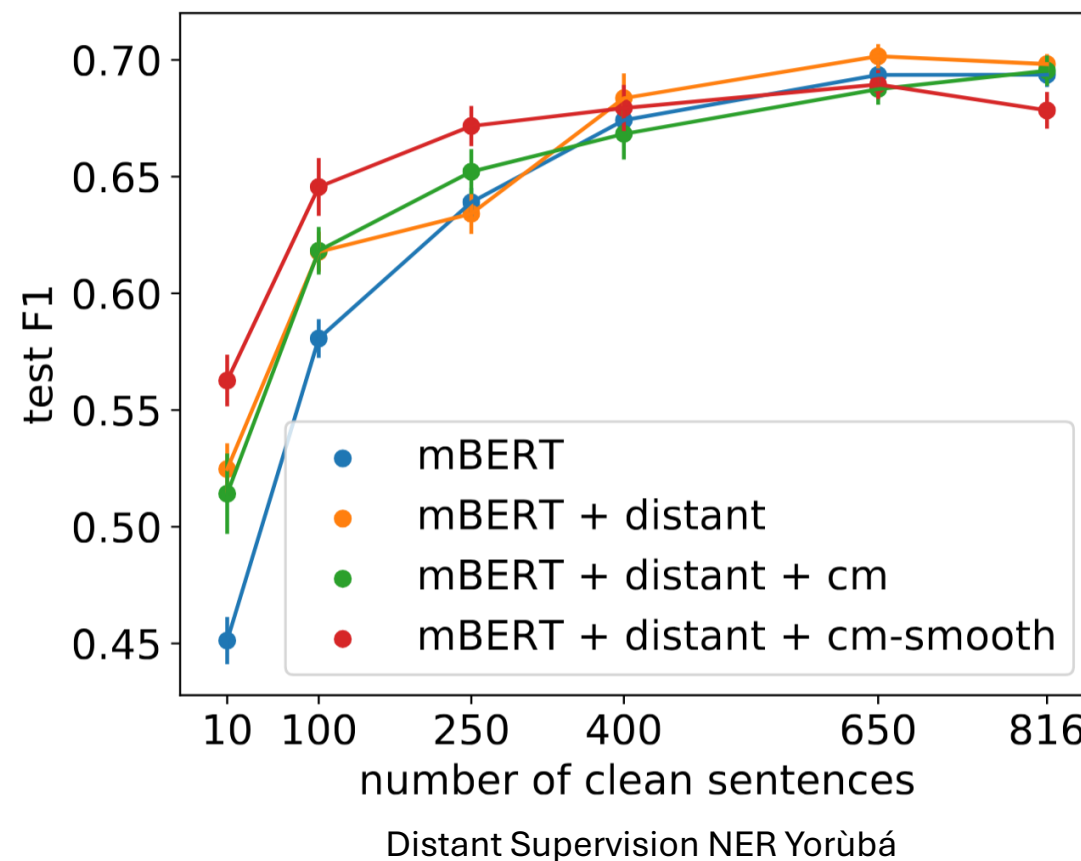
<https://gratisography.com/photo/lonely-road-blue-sky/>
<https://unsplash.com/de/fotos/luftaufnahme-der-kurvenreichen-strasse-6aDwXoWeElc>

Once upon a time...

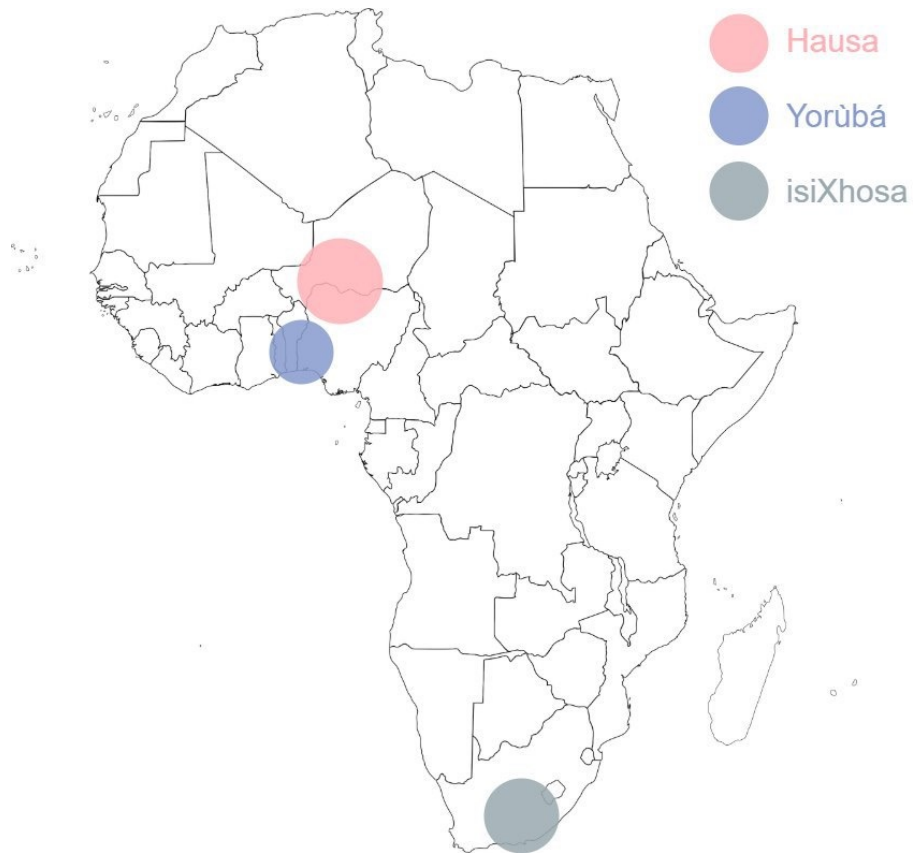


Low-Resource NLP

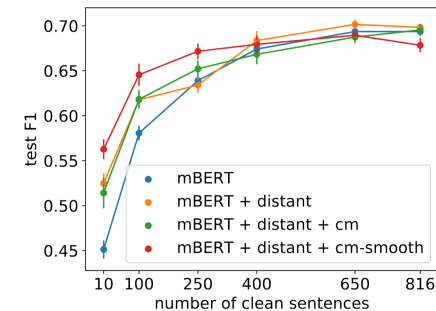
Multilingual BERT
+ Distant/Weak Supervision
(automatic data annotation)



Low-Resource Techniques for African Languages



- Widely spoken and digital presence
- Few NLP datasets or tools



Hedderich, Adelani, Zhu, Alabi, Markus & Klakow:
Transfer Learning and Distant Supervision for Multilingual Transformer Models: A Study on African Languages (EMNLP'20)

Low-Resource Techniques for African Languages

- In this project: Collaboration with/access to local language speakers, annotators & programmers
- Setting up low-resource techniques vs. just annotating a bit more

	Weak supervision	Manual annotation	
NER	ca. 30 minutes with ANEA (our toolkit)	ca. 30 sentences	😊
Topic classification	ca. 2.5 hours with programming	ca. 1500 sentences	🤔

The Limits of Benchmarks and Datasets



Assumptions

- Benchmarks / datasets are great
 - Quantitative, reproducible, fast evaluation, ...
- Benchmarks/datasets take a simplified view of the world
 - For low-resource techniques
 - Assumptions on expertise (coding + domain), existence of support tools, ...
 - Only if assumptions were met, the benchmarks provided useful guidance
 - Single true label assumption
 - Ignores diversity, subjectivity and disagreement
 - Barbara Plank: “The 'Problem' of Human Label Variation” (EMNLP’22)
 - LLM-as-a-Judge & LLM-generated datasets
 - Assumes LLM is a faithful representation of real humans
 - Annemarie Friedrich: Conference Report ACL 2025

The Limits of Benchmarks and Datasets

How can we uncover unknown assumptions
and test known assumptions?

Uncovering and Testing Assumptions

Ask real users*!

* user $\in$ {local language speakers, domain experts, developers, NLP engineer, end user, ...}

How do you ask real users scientifically?

Human-Computer
Interaction



Computer scientists, psychologists,
designers, social scientists, ...

- **Methodology:** quantitative + qualitative user studies, exploratory interviews, demos, ... *
 - **Interaction:** Not just designing pretty interfaces
 - **Mindset:** human-focused and future thinking
- * Issues of their own (reproducibility, transparency, technical implementations, ...)

Menu of the Day

- HCI Project
Supporting Teachers in Controlling
Chatbots
- NLP Project
Supporting Users in Understanding Effects
of Prompt Changes
- Bigger Picture & Discussion

Menu of the Day

- HCI Project
Supporting Teachers in Controlling
Chatbots
- NLP Project
Supporting Users in Understanding Effects
of Prompt Changes
- Bigger Picture & Discussion

A Piece of Theatre: Investigating How Teachers Design LLM Chatbots to Assist Adolescent Cyberbullying Education

Michael A. Hedderich

Natalie N. Bazarova

Wenting Zou

Ryun Shim

Xinda Ma

Qian Yang



Chatbots in Education

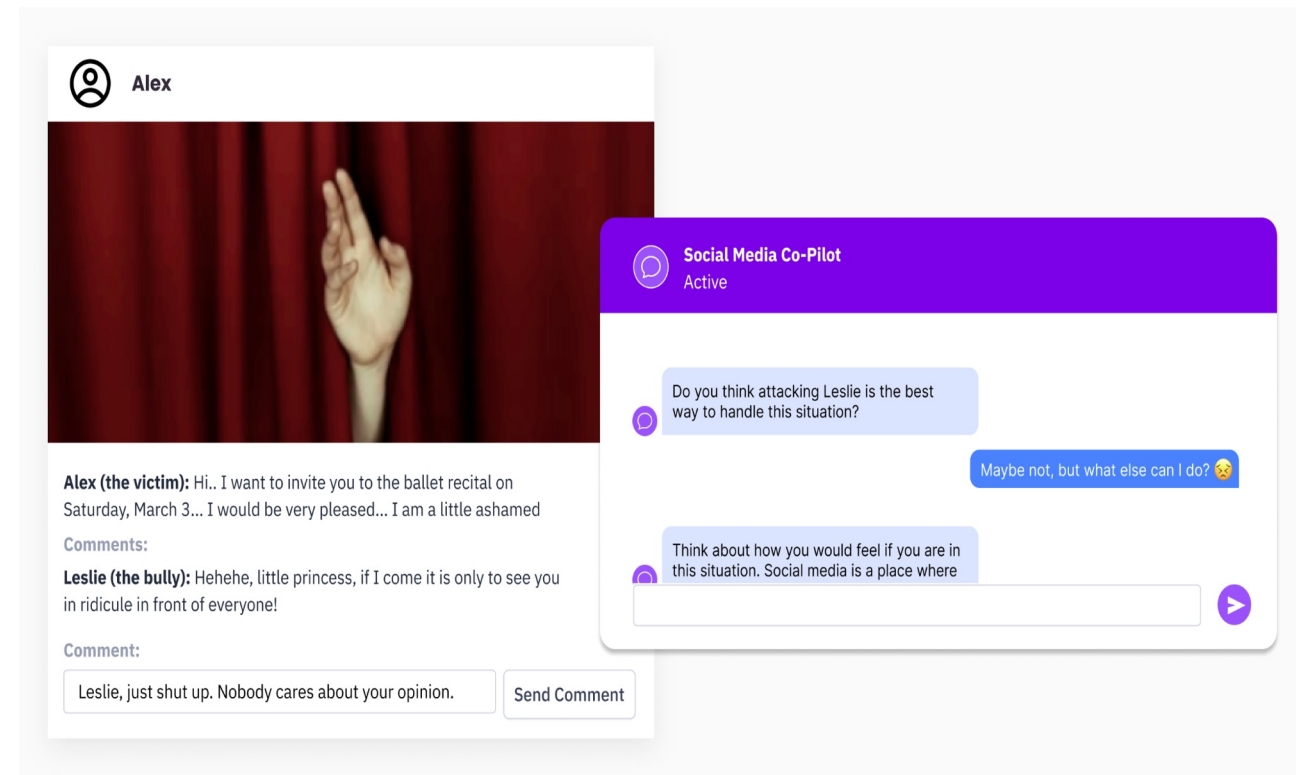
- Chatbots as support tools for teachers in class
 - Immediate feedback
 - Personalized attention

“

[The chatbot] keeps those wheels turning.

~ Teacher #2

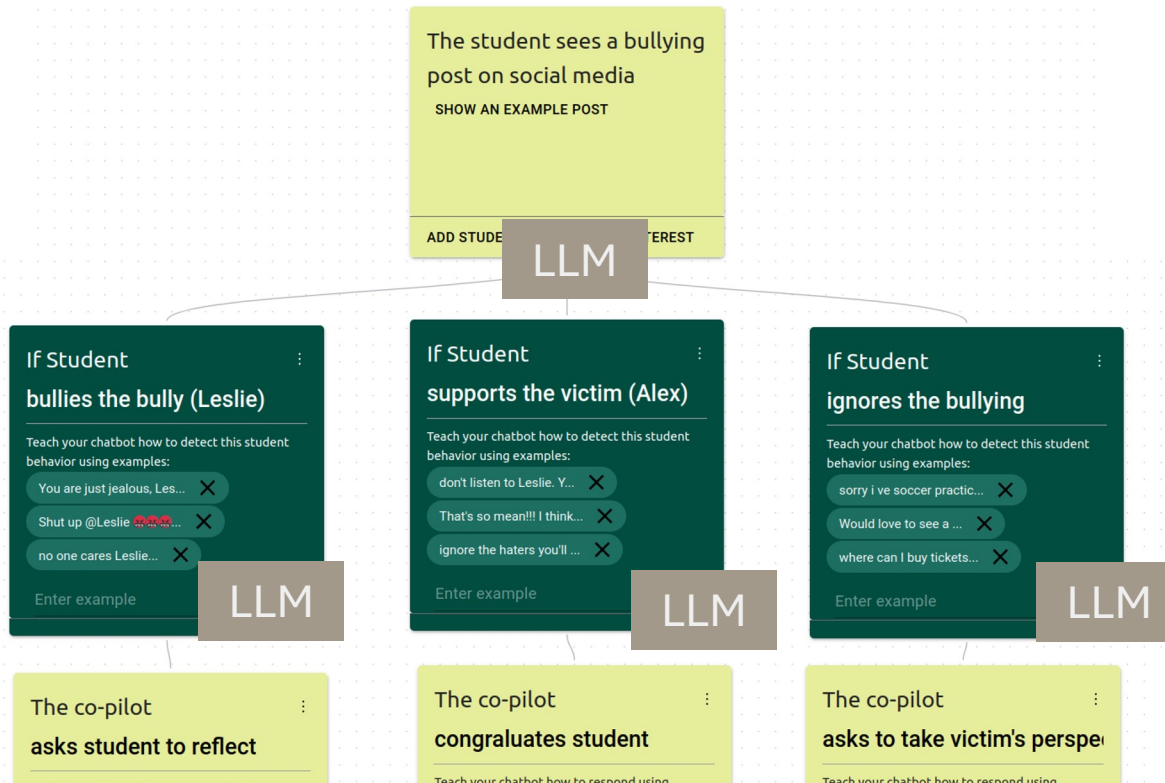
SocialMediaTestdrive.org on Cyberbullying



Combining Dialogue Structure + LLMs

Social Media Co-Pilot

🔍 Fake News Example 📄 New Cyberbullying Design 🔍 Load Design



- Teachers: Specify structure + examples
- Backend: LLM few shot classifiers + generators

User Study

- Middle school teachers build and test their own chatbot
 - For interacting with a student about cyberbullying
 - Test chatbots interactions
 - Gain experience
- Deeper insights than just discussing hypothetical situations
- Collect teachers' experiences through think-aloud and semi-structured interviews

System as a
Probe

Insights Into Users & Needs for NLP Technology

- Not just a teacher feedback replacement
 - Roleplays with chatbots (students try out responses in safe environment)
 - Teacher as a playwright steering and allowing for improvisation
- Trading off flexibility and control for complex tasks
 - Real users struggle with controlling LLMs (e.g. J.D. Zamfirescu: “Why Johnny Can’t Prompt”/ “Herding AI Cats”)
 - LLM-Chains gave more control on the fine-grained level
 - LLM component improvising for the specific student situation
 - Limited guidance on more abstract level (dwelling on a topic, jumping to a different topic, controllable guardrails and fixpoints)
- Collaborative chatbot design
 - Changes on different levels (developer, school district, teacher, ...)
 - Ways to adapt models (finetuning, representation engineering/model edits, prompting)

Validate NLP work +
new NLP challenges

Menu of the Day

- HCI Project
Supporting Teachers in Controlling
Chatbots
- NLP Project
Supporting Users in Understanding Effects
of Prompt Changes
- Bigger Picture & Discussion

A Common Struggle

Application partners

Archaeologists, designers, educators,
medical researchers, ...

Excited to try out LLMs for their
tasks

Low-entrance barrier through prompting

Struggle to evaluate and understand the effects of prompt
changes (prompt engineering)



Picture by 2H Media auf Unsplash

Prompt Engineering Challenge

Tell a short story about a day in the life of the doctor Dr. Li.



Dr. Li is a young and ambitious doctor who has just opened up his own clinic in a small town ...

Dr. Li is a general practitioner who has been practicing medicine for over 20 years and has seen it all ...

Dr. Li is a general practitioner in a small town in the countryside. She has been practicing medicine for over 20 years and has seen it all ...

280 male: 220 female

Tell a short story about a day in the life of the doctor Dr. Smith.



Dr. Smith is a general practitioner who has been working in the small town of Willow Creek for over 20 years ...

Dr. Smith is a general practitioner who has been working in the small town of Willow Creek for over 20 years ...

Dr. Smith was a general practitioner who had been in practice for over 20 years ...

499 male: 1 female

Gender Ratio

Llama 2 examples

Existing Evaluation Methods

Automatic Metrics



- Accuracy, BLEU, special classifiers
- Fast and cheap
- Predefined criteria or data
- Limited insights

Full Human Evaluation

Texts
Dr. Li is a young and ambitious doctor who has...
Dr. Li was a brilliant and compassionate ...
Dr. Smith is a general practitioner who has ...
Dr. Smith was a busy man. He has practice in ...

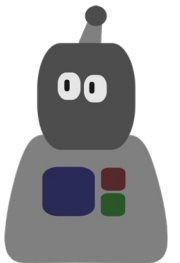


- Flexible + insightful
- Effort
- Biases, inconsistency, false patterns

LLM as Judge

- Fast and mostly cheap
- Biases, stability, lack of correlation with humans

Middle Ground?



The Spotlight Framework

**Label-Descriptive Patterns and Their Application to
Characterizing Classification Errors**

**What's the Difference? Supporting Users in Identifying the Effects of
Prompt and Model Changes Through Token Patterns**



Jonas
Fischer



Jilles
Vreeken



Dietrich
Klakow



Barbara
Plank



Anyi Wang



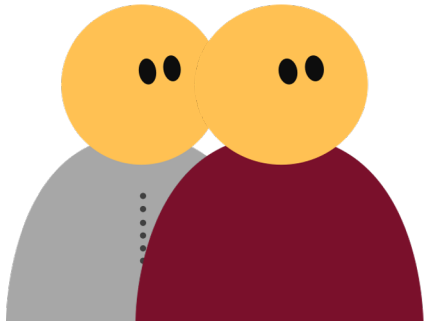
Raoyuan
Zhao



Florian
Eichin

Understanding Effects of Prompt / Model Changes

- LLM outputs are long and nuanced texts
- Even small prompt or model changes can have large effects
- Support users in better understanding effects of their changes
 - End users: guide prompting → better results
 - Researchers: guide model inspection → better understanding of model behavior



25.09.2025

Michael A. He
Methods in Human-

Dr. Smith was a general practitioner who had been in practice over 20 years. He had a loyal patient base and was well respected in the community. Here is a short story about a typical day in the life of Dr. Smith:\n\nDr. Smith's day began early, as he always made it a point to be at the clinic by 8am. He started adjusting the patient schedule for the day and making any necessary adjustments. He also spent some time checking on his first patient who was running smoothly for several days. After the thorough examination, he prescribed some medication and discussed the patient's progress.

Finding Systematic Differences

- Challenge:
 - Even with same LLM and prompt, outputs are not identical (stochastic decoding)
 - Systematic and consistent differences between two groups of LLM outputs (two prompts or two models with same prompt)

Dr. Li is a young and ambitious doctor who has just opened up his own clinic in a small town ...

Dr. Smith is a general practitioner who has been working in the small town of Willow Creek for over 20 years ...

Dr. Li was a brilliant and compassionate doctor who had been practicing medicine for over 20 years. She was known for her dedication to her patients ...

Dr. Smith was a busy man. He had a practice in a small town, where he was the only doctor for miles around. He saw patients of all ages, ...

Finding Systematic Differences

- Challenge:
 - Even with same LLM and prompt, outputs are not identical (stochastic decoding)
 - Systematic and consistent differences between two groups of LLM outputs (two prompts or two models with same prompt)
- Spotlight Framework: Automation + Human Analysis
 - Automatically describe the systematic differences between two groups of LLM outputs
 - Token patterns using data mining techniques
 - Human analyzes these (unexpected) differences

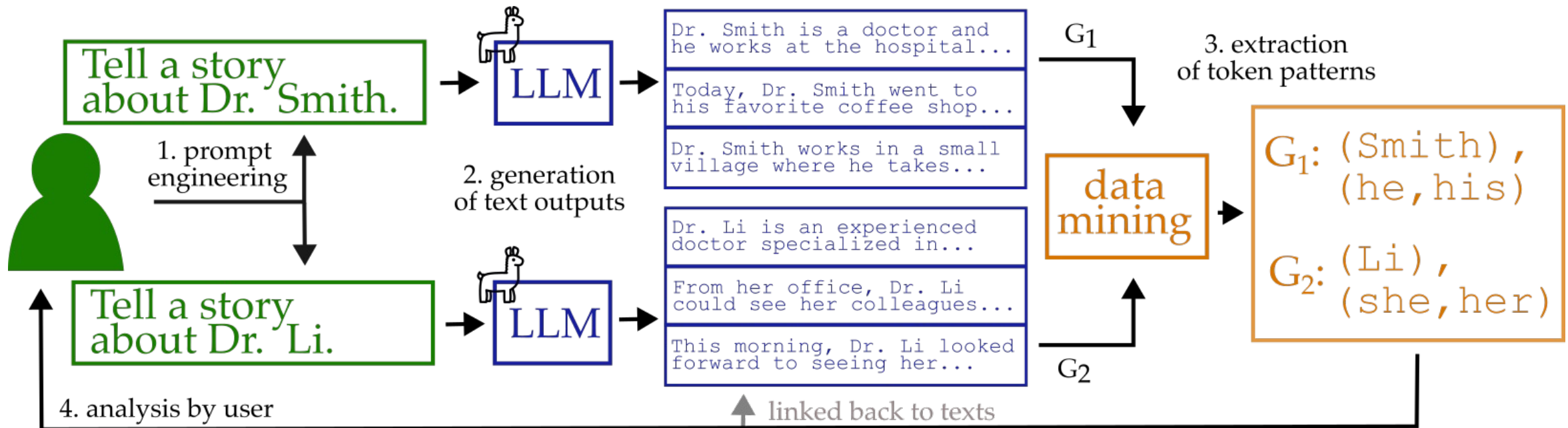
Tell a short story about a day in the life of the doctor Dr. Li.



Tell a short story about a day in the life of the doctor Dr. Smith.

(He, he, his)

Our Spotlight Approach



Data Mining

Technical Challenges

- Scaling to complexity of textual data
- Resistance to noise in natural text

$$L(\pi_I(D) | I) = \log \left(\frac{|D|}{|\sigma_I(D)|} \right)$$

$$L(\pi_I(D) | I) = \log \left(\frac{|D^-|}{|\sigma_I(D^-)|} \right) + \log \left(\frac{|D^+|}{|\sigma_I(D^+)|} \right)$$

$$L(M) = L_{\mathbb{N}}(|\mathcal{P}|) + \sum_{P \in \mathcal{P}} (L_{\mathbb{N}}(|\gamma(P)|) + L_{pc}(|D^+|) + L_{pc}(|D^-|)) + \sum_{cl \in \mathcal{P}} \left(\log \left(\frac{|\mathcal{I}|}{|cl|} \right) + L_{pc}(|\mathcal{I}|) \right) + \sum_{I \in \mathcal{I}} L_{pc}(|D|)$$

Hedderich, Fischer et al.: Label-Descriptive Patterns and Their Application to Characterizing Classification Errors (ICML 2022)

Evaluation of Spotlight

3 Benchmarks



**Existing prompt
data**



All details and results
in the paper!

**Demonstration
studies**



User study



Demonstration Study – Model Change

Tell a 50 word story about a farmer from Kenya.



GPT 3.5-turbo

(Mwai, Mwai's)

(Wanjiku,
Wanjiku's)

...

**Local Name
Distribution**

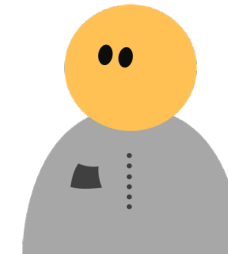
Medium + Low
Frequency Names

One name
dominates (>60%)



GPT 4o-mini

(Juma)



User Study

- Generate two groups of texts (from same source)
- Insert differences (known ground truth)
- Crowd worker study
- Simplified setting, no mention of AI, small texts

season's labor and nature's abundance on the land.

After months of hard work, the farmer gathers ripe crops - golden maize and vibrant vegetables - under a bright sky, celebrating the season's bounty and preparing for the next planting cycle.

Under the golden sun, a farmer gathers vibrant crops, filling baskets with ripe maize and beans, and lush vegetables celebrating nature's bounty while preparing for market and ensuring sustainable practices for future seasons.

The farmer eagerly gathers ripe corn and tomatoes, sunlight warming the fields. Lush produce fills baskets, ready for market. A successful season brings hope and nourishment to the community.

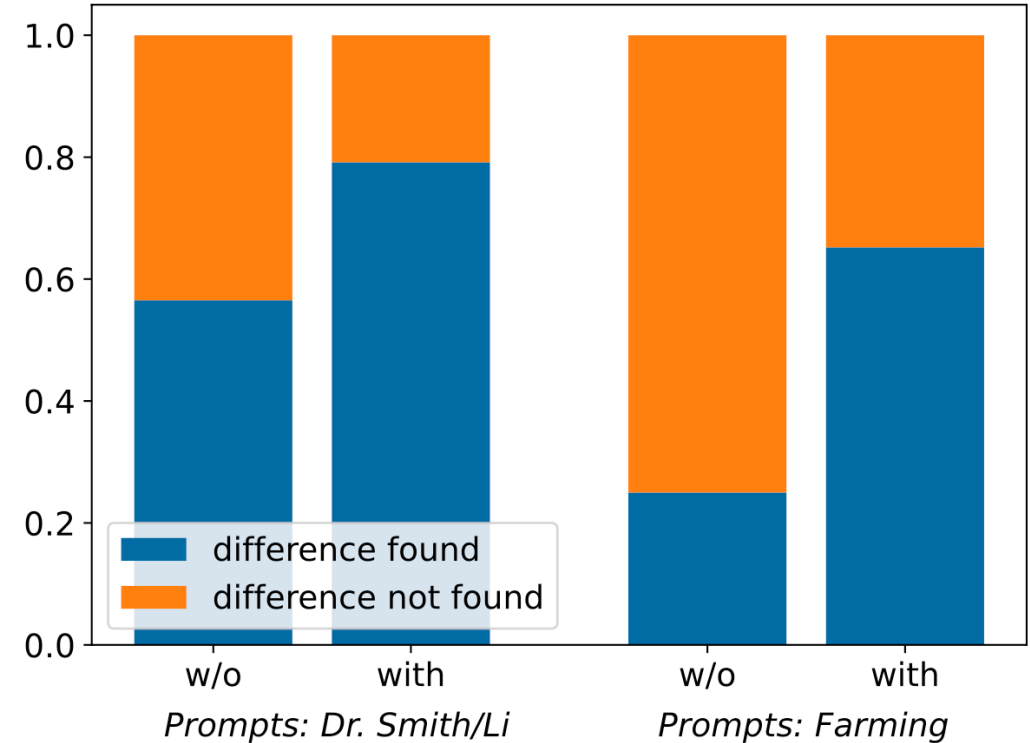
The farmer gathers ripe crops from sprawling fields, using machinery to efficiently collect grains and vegetables. Sunlight and hard work transform the land into bountiful yields, ensuring a prosperous season.

The farmer diligently gathers golden grains and vibrant fruits, working alongside machines and the sun. Each bushel represents hard work, hope, and the cycle of life, ready for market.

The following word(s) occur(s) differently between the two groups of text:
maize, beans (more often in group 1)

Does It Actually Help Users

- Yes!
- Assumption holds:
LLM output differences hard to detect for users
- Users identify differences that are not systematic (false positives)
- Very simplistic user study
 - Realistic settings
 - Part of prompt engineering process
 - Gero et al.: “Supporting Sensemaking of Large Language Model Outputs at Scale” (CHI’24)



Explaining Differences Between Groups of Text

 README  MIT license  

PyPremise

[Premise](#) is a data mining method to identify significant differences between two groups of texts. It can be used, e.g., to identify from the outputs of an LLM what systematic differences occur if you change your prompt. Or it can identify patterns or explanations of where an NLP classifier performs well and where it fails.

PyPremise is a Python library to makes it easy to run Premise with just a couple of lines of code.

Example Applications

Systematic differences between two prompts or two LLMs

Premise identified, e.g., that changing the last name of the protagonist from Dr. Li to Dr. Smith in a story generation prompt caused - among other differences - a drastic change in the gender distribution (more details and examples [here](#))

Patterns Group 0 (Dr. Li)	Patterns Group 1 (Dr. Smith)
(She, she)	(He, he)
(car, accident)	(general, practioner)
	(small, town)



PyPremise

```
pip install pypremise
```


Menu of the Day

- HCI Project
Supporting Teachers in Controlling
Chatbots
- NLP Project
Supporting Users in Understanding Effects
of Prompt Changes
- Bigger Picture & Discussion

Behavior of Black Box LLM/AI

Understanding

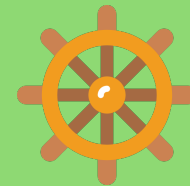
Behavioral
Analysis



Interpretability



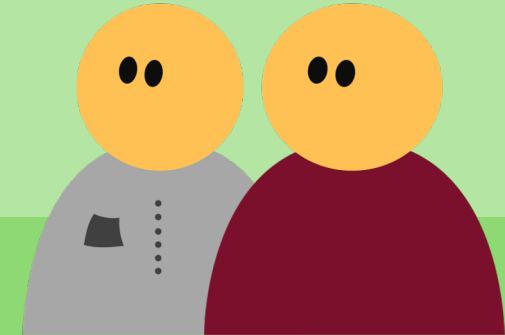
Control



ML/NLP

HCI

- ← User needs
- User evaluation
- Applications



What is Human-Centric NLP?

Human-Centric Methods

to study NLP tasks and methods in human context

Human-Centric Tasks

e.g. Human-AI Collaboration

Beyond just more (human-centric) benchmarks

Who Should Perform Human-Centric NLP?

At NLP venues

- Getting more open to HCI methods?
- Is it the right venue?

At HCI venues

- Experts in human-centric methodology
- Already studying and using LLMs
- Challenging for NLP people to enter
- NLP can bring in useful expertise

Together

- Establish shared vocabulary/understanding
- Open a whole new world of interesting insights

Other Perspectives



**Annemarie
Friedrich**

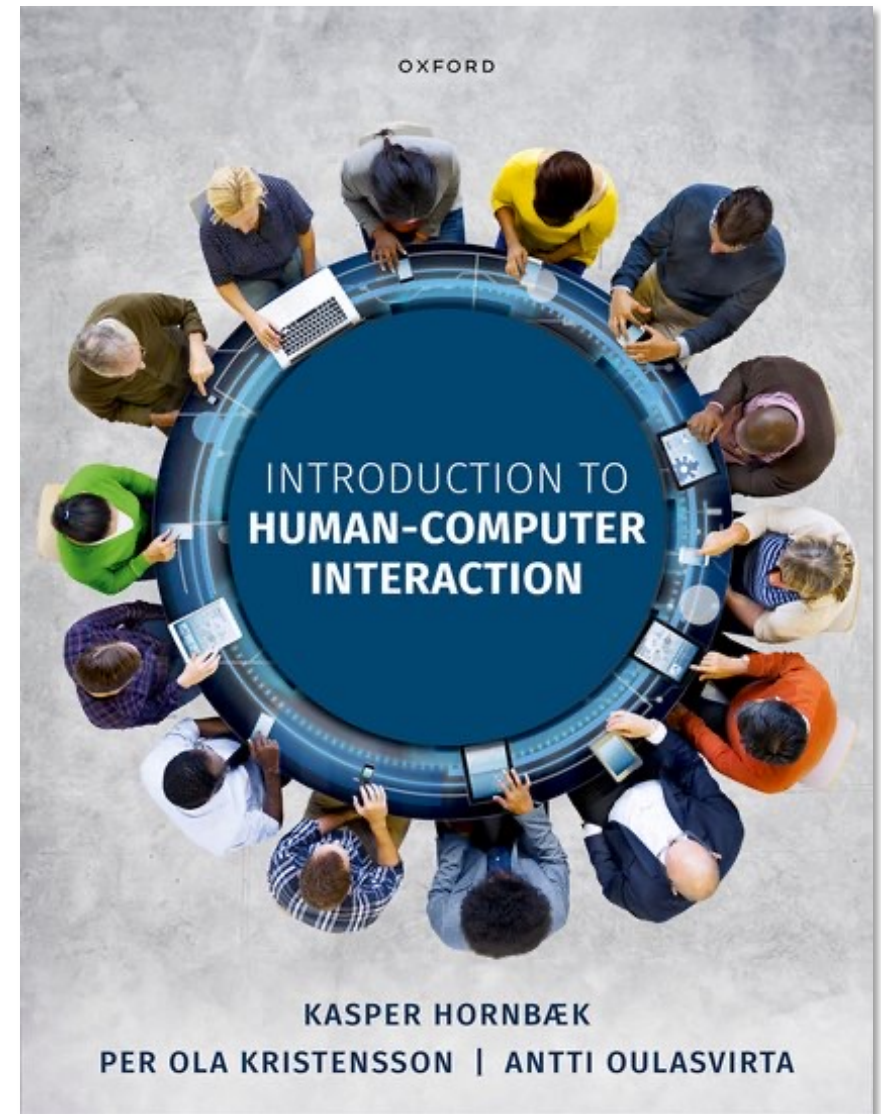
Computational
Linguist

Conference Report ACL 2025 - How can we move forward as a field?

30 minute read

 **Published:** August 02, 2025

DISCLAIMER: If you disagree with me, or if I represented your opinions wrongly in this article, please contact me directly via e-mail! I am happy to adapt this article if needed.

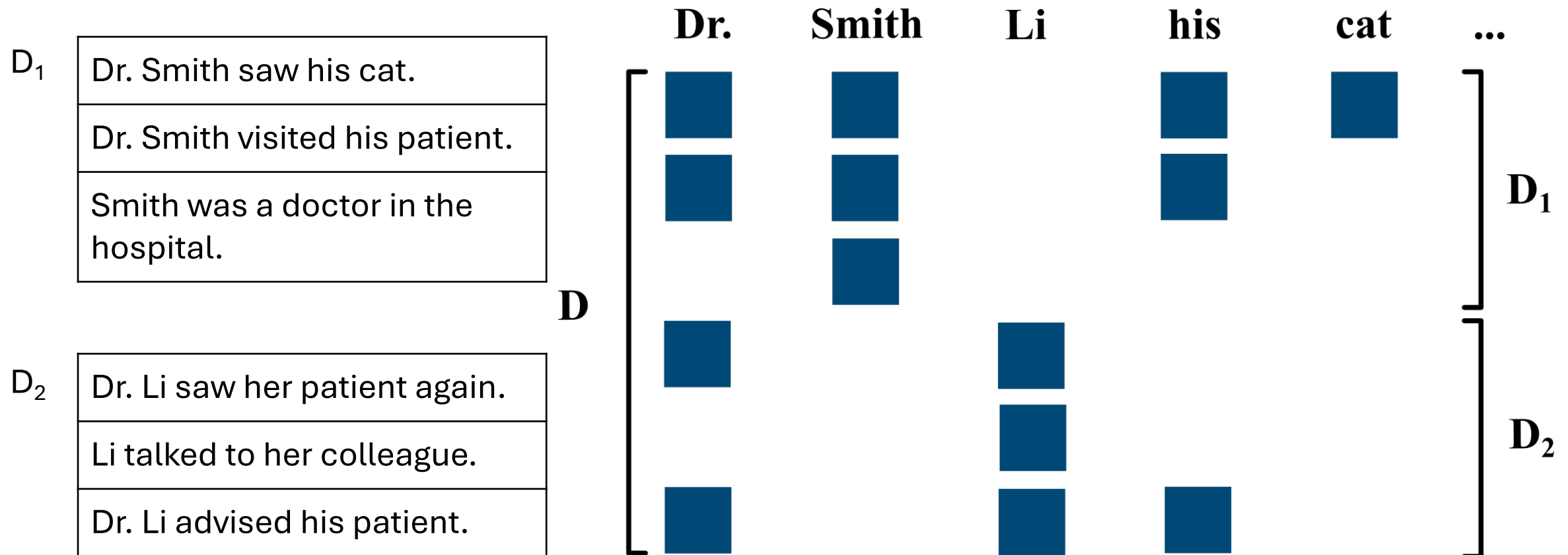


Thank you!

- Benchmarks, datasets and LLM-as-a-Judge are based on **assumptions**
- **Human-centric methods** can validate our **known** and uncover **unknown** assumptions
- **HCI methodology**
 - Probes with semi-structured interviews
 - Demonstration & user studies
- How should NLP **collaborate** with HCI?



Leveraging Data Mining Techniques



The Patterns We Are Looking For

Searching for token patterns
that show probability shift
between D_1 and D_2

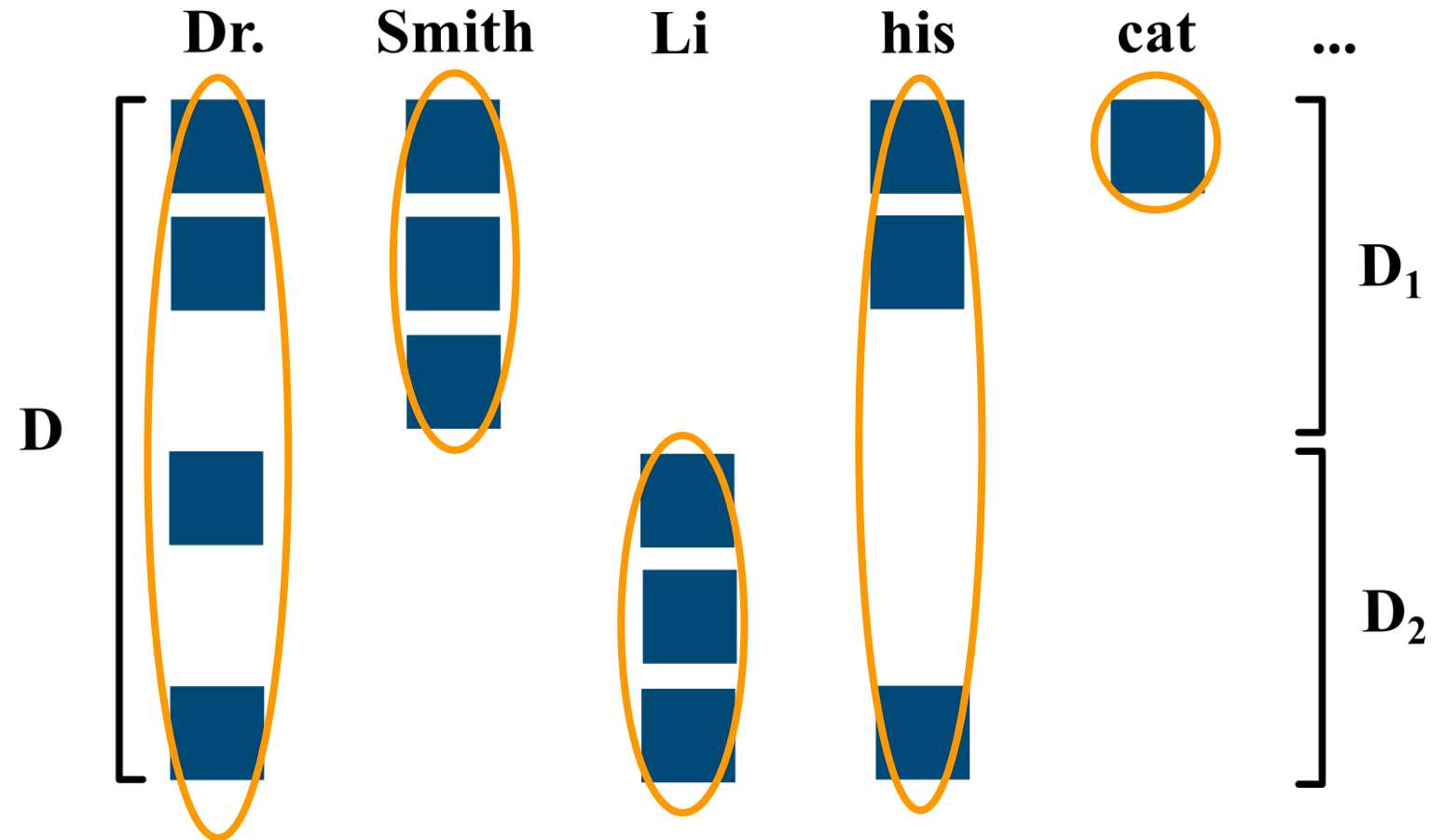
Dr. 🙄

Smith 👍

Li 👍

his 🤔

cat 🤔

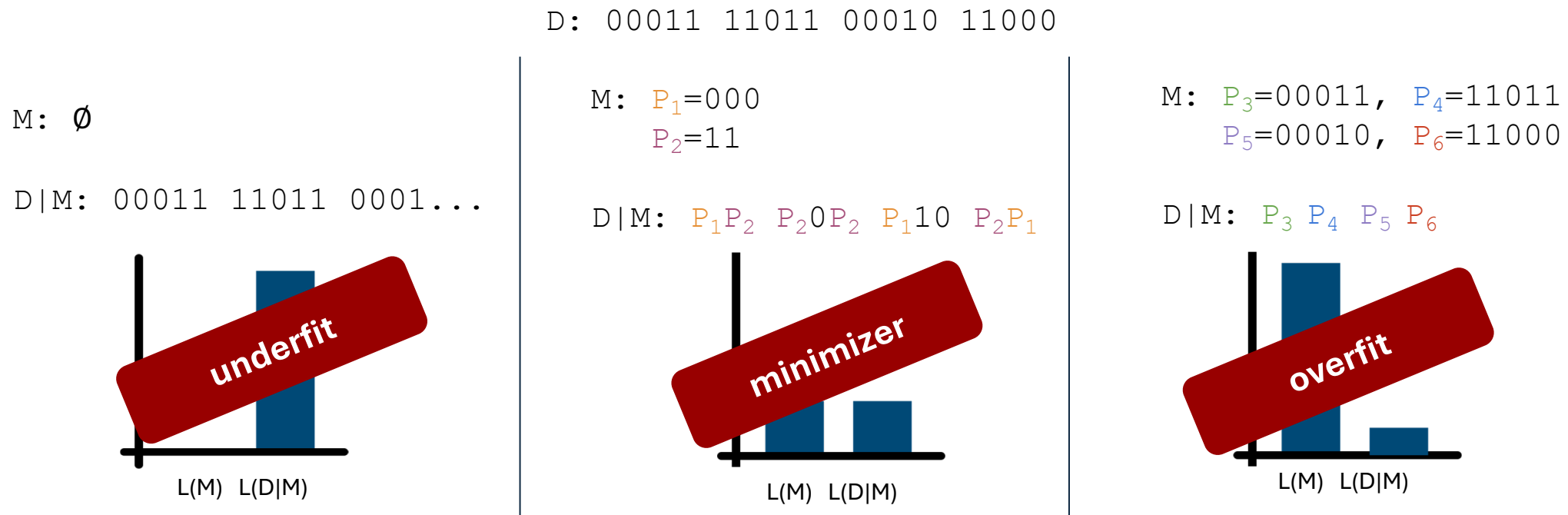


What is Relevant and What Not?

- Distinguish between systematic differences and random noise
- Minimum Description Length principle
 - without held-out data
 - based on loss-less compression

A Minimal Intro to Minimum Description Length (MDL)

- Robust way to find systematic patterns
- Model M (patterns) that obtains best lossless compression of data D
- Minimize length $L(M) + L(D|M)$



Spotlight in Action – Prompt Change

Tell a short story about a day in the life of a doctor.



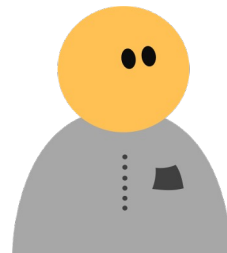
(She, she)

(car, accident)

Gender bias

Occupation

Location



Tell a short story about a day in the life of the doctor Dr. Smith.



(Smith)

(He, he)

(general,
practitioner)

(small, town)